

# Knowledge Evolution in Physics Research: An Analysis of Bibliographic Coupling Networks

Wenyuan Liu (刘文源),<sup>1</sup> Andrea Nanetti,<sup>2</sup> and Siew Ann Cheong<sup>1, \*</sup>

<sup>1</sup>*School of Physical and Mathematical Sciences,  
Nanyang Technological University,  
Singapore 637371, Singapore*

<sup>2</sup>*School of Art, Design and Media,  
Nanyang Technological University,  
Singapore 637371, Singapore*

(Dated: April 5, 2017)

Even as we advance the frontiers of physics knowledge, our understanding of how this knowledge evolves remains at the descriptive levels of Popper and Kuhn. Using the APS publications data sets, we ask in this letter how new knowledge is built upon old knowledge. We do so by constructing year-to-year bibliographic coupling networks, and identify in them validated communities that represent different research fields. We then visualize their evolutionary relationships in the form of alluvial diagrams, and show how they remain intact through APS journal splits. Quantitatively, we see that most fields undergo weak Popperian mixing, and it is rare for a field to remain isolated/undergo strong mixing. The sizes of fields obey a simple linear growth with recombination. We can also reliably predict the merging between two fields, but not for the considerably more complex splitting. Finally, we report a case study of two fields that underwent repeated merging and splitting around 1995, and how these Kuhnian events are correlated with breakthroughs on BEC, quantum teleportation, and slow light. This impact showed up quantitatively in the citations of the BEC field as a larger proportion of references from during and shortly after these events.

According to Karl Popper, science progresses through repeated hypothesis testing [1]. Hypotheses contrary to empirical evidence must be rejected, while those consistent with data survive to be tested another day. In this picture of the scientific enterprise, our knowledge of the world around us is always tentative, but becomes more complete over time. On the other hand, Thomas Kuhn believes that the accepted knowledge of a given time is the result of consensus amongst scientists, based on evidences consistent with their theories [2]. However, when too many conflicting evidences are found, a new consensus can form around new theories in what he called a “paradigm shift”. Kuhn gives special relativity and quantum theory as examples of paradigm shifts. Looking back, we realize these two theories have enormous impacts on how we understand the world today. But could there be paradigm shifts of various scales that have also contributed to reshaping our knowledge of physics?

Many historians of science have noted the strongly reductionistic flavor of scientific research in the last couple of centuries[3]. Starting as natural philosophy, the body of scientific knowledge became separated disciplines of astronomy, biology, chemistry and physics. Within physics itself, we also observe the emergence of high energy physics, condensed matter physics, biophysics, and photonics. These are the results of the splitting of science into more specialized fields. We also observe in parallel the merging of fields, such as the merging of astronomy and physics to give astrophysics, biology and chemistry to give biochemistry, and others “that arose

by division and recombination of specialties already matured” [2]. These developments have been discussed extensively by philosophers and historians of science, but unlike our quantitative understanding of physics, our appreciation for the processes through which we acquired our knowledge of physics remains at a highly descriptive level. Some progress has been made in addressing this problem [4–6]. In particular, the following three papers provide the inspiration for our study. Chen and Redner suggested that long-range connections can form between disparate fields because of the development of “a widely applicable theoretical technique, or cross fertilization between theory and experiment” [7]. Visualizing the cross citations between neuroscience journals, Rosvall and Bergstrom traced the growth and maturation of neuroscience as a discipline [8]. Using embryology as a specific example, Chavalarias and Cointet created a phylomemetic network visualization for the evolution of science [9].

While these previous studies point to the evolution of scientific knowledge, they do not identify the entity that is recognizably ‘knowledge’, or they do not study the interactions between such objects. To clarify what constitutes knowledge, we start with the bibliographic coupling network (BCN) [10], proposed by Kessler and used extensively in computer science [11, 12]. In a BCN, nodes represent papers, and if two papers share  $w$  common references, we draw an edge with weight  $w$  between them (see Fig. 1(a)). The BCN is suitable for our purpose for two reasons: (i) the BCN for a given year consists only of papers published that year and does not change after more papers are published later, so features in the BCN represent the state of knowledge in that year; and (ii)

\* cheongsa@ntu.edu.sg

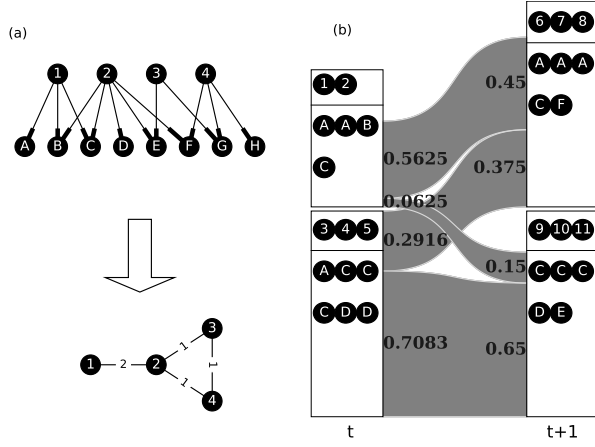


FIG. 1. (a) Building a BCN (lower) from a citation network (upper): circles with numbers are papers under consideration, circles with letters are their references, and numbers on edges are weights. (b) Topical clusters in year  $t$  (left) and in year  $t + 1$  (right), and their forward (left) and backward (right) intimacy indices, shown as flows.

the appropriate collective unit of knowledge is a community in the BCN instead of a few key papers or a journal. For the American Physical Society (APS) data set, consisting of about half a million publications between 1893 and 2013 [13], we show in Supplementary Information SI Sec. I that the BCN edge weights are far more heterogeneous than expected from an appropriate null model.

This heterogeneity can be explained by the presence of communities that we extracted using the modularity-optimized Louvain method [14]. Using the null model in SI Sec. I, we show that these communities are statistically significant. We also test in SI Sec. II how likely the most common PACS number in a community of  $n$  papers can appear with its observed frequency, within random collections of  $n$  papers. For most communities, this is highly unlikely, so we conclude that the groupings of papers extracted are meaningful. We refer to these validated units of knowledge as *topical clusters* (TCs).

To study how knowledge evolves, we investigate how TCs  $\{C^t\}$  in year  $t$  become  $\{C^{t+1}\}$  in year  $t + 1$ . The papers published in different years are distinct, but they do overlap in their references. Therefore we use this fact to define a *forward intimacy index*  $I_{mn}^f$  and a *backward intimacy index*  $I_{mn}^b$ :

$$I_{mn}^f = \sum_i \frac{N(R_i, \mathcal{R}_n^{t+1})}{N(R_i, \mathcal{R}^{t+1})} N(R_i, \mathcal{R}_m^t) / L(\mathcal{R}_m^t);$$

$$I_{mn}^b = \sum_i \frac{N(R_i, \mathcal{R}_m^t)}{N(R_i, \mathcal{R}^t)} N(R_i, \mathcal{R}_n^{t+1}) / L(\mathcal{R}_n^{t+1}),$$
(1)

to quantify how close  $C_m^t$  is to  $C_n^{t+1}$ . Here the TCs at  $t$  and  $t + 1$  are  $C^t = \{C_1^t, \dots, C_m^t, \dots, C_u^t\}$  and  $C^{t+1} = \{C_1^{t+1}, \dots, C_n^{t+1}, \dots, C_v^{t+1}\}$ , and we denote the references

cited by papers in  $C_m^t$  and  $C_n^{t+1}$  as  $\mathcal{R}_m^t = \mathcal{R}(C_m^t) = [R_1^m, \dots, R_p^m]$  and  $\mathcal{R}_n^{t+1} = \mathcal{R}(C_n^{t+1}) = [R_1^n, \dots, R_q^n]$ ; and  $\mathcal{R}^t = \{\mathcal{R}_1^t, \dots, \mathcal{R}_m^t, \dots\}$ .  $N(\text{element}, \text{list})$  is the number of times *element* occurs in *list*, and  $L(\text{list})$  is the length of *list*. In this definition, we assume each citation instance in  $t$  will be uniformly distributed over all instances of the same citation in  $t + 1$ , while each citation in  $t + 1$  receives equal contributions from all instances of the same citation in  $t$ . In general, this index is asymmetric, i.e.  $I_{mn}^f \neq I_{mn}^b$ , because the references are not cited the same number of times in the two years, as illustrated in Fig. 1(b).

We visualize the sequence of TCs and their intimacy indices, the evolution of physics research they represent in the form of alluvial diagrams. For example, in Fig. 2 we can clearly see the birth of PRA, PRB, PRC and PRD from PR in 1970. Each journal consist of several TCs, which existed even in the PR era. The editorial decision to split PR is consistent with the self-organized TCs even though it was done without classification analysis. We also see the consistent birth of PRE from PRA in 1993 In SI Sec. III.

More importantly, from the alluvial diagram we can identify the key interactions between TCs that are correlated with important publications. In Fig. S4 of SI Sec. IV we showcase one such episode between 1991 and 2000. At the beginning of the decade, we see two PRA-dominated TCs. Based on the papers they contain, we can loosely associate one with quantum information (QI) and trapped atomic ions (BEC), and the other with quantum optics (QO). In 1993, the QI + BEC TC cited many QO papers, and in 1994, the QO TC cited many QI + BEC papers. Following this ‘cross-fertilization’, the two TCs merged in 1995, the same year Cornell *et al.* [15] and Ketterle *et al.* [16] published their seminal papers demonstrating BEC in dilute atomic gases. In recognition of their works, Cornell, Wieman, and Ketterle were awarded the 2001 Nobel Prize in Physics. The PRA-dominated TC split after 1996 to give one that is exclusively BEC, and another that is still a combination of QI + QO. It was after Zeilinger demonstrated in 1997 experimental quantum teleportation [17] that the QI + QO TC split into a QI TC and a QO TC. After receiving more influence from other PRB-dominated TC, the QO cluster produced yet another breakthrough paper, in the form of ultraslow light in hot atomic gases[18]. Without the data visualization done here, few may suspect the existence of such connections between BEC, quantum teleportation and slow light.

From Fig. 2 we see a diversity of inflows and outflows from one TC to another: some TCs are derived almost exclusively from one source, others receive strong contributions from a small number of sources, or weak contributions from a large number of sources. To quantify such diversity, we construct a forward mixing degree of community  $C_m^t$  and backward mixing degree of  $C_n^{t+1}$  analo-

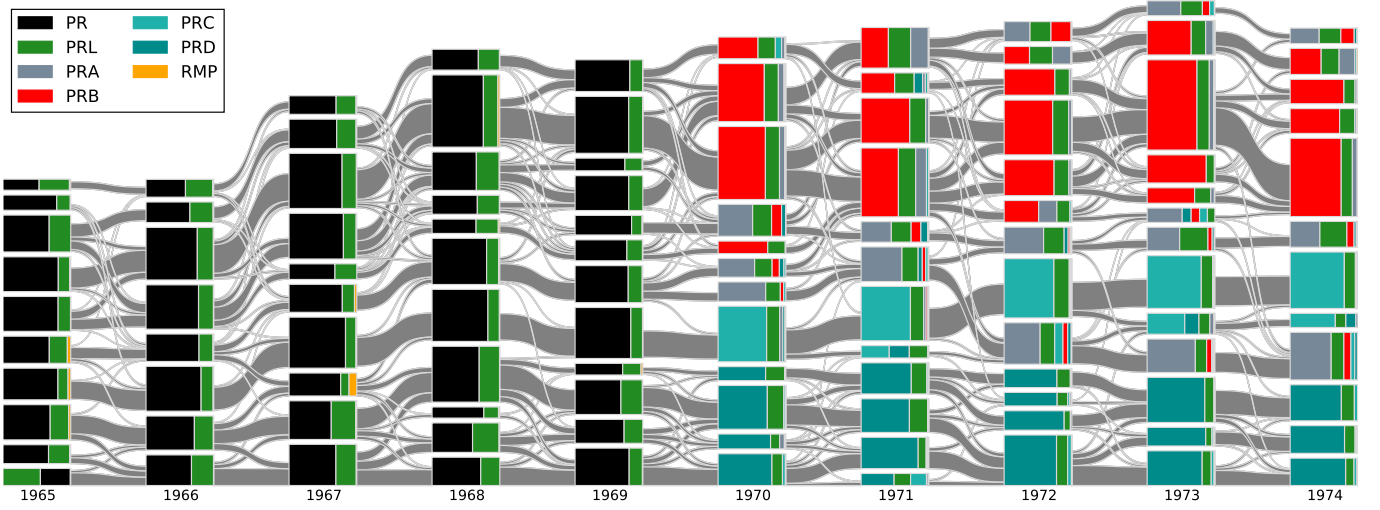


FIG. 2. The alluvial diagram of APS papers from 1965 to 1974. Each block in a column represents a TC and the height of the block is proportional to the number of papers in the TC. Only communities comprising more than 100 papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the forward and backward intimacy indices. The different colors in a TC represent the relative contributions from different journals.

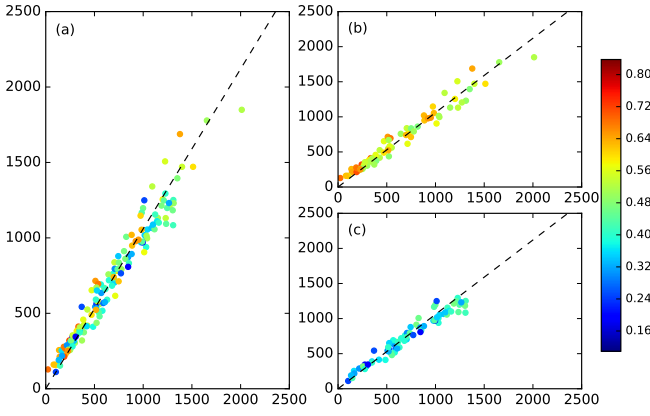


FIG. 3. (a) Plot of observed (y-axis) against predicted (x-axis) sizes of recombined TCs, showing a linear growth with slope 1.06 (dashed line). This linear growth is the same for TCs with (b) high (red) or (c) low (blue) backward mixing degree.

gous to the Gini-Simpson index [19]:

$$\begin{aligned} M_m^f &= 1 - \sum_n \left( I_{mn}^f / \sum_{n'} I_{mn'}^f \right)^2, \\ M_n^b &= 1 - \sum_m \left( I_{mn}^b / \sum_{m'} I_{m'n}^b \right)^2. \end{aligned} \quad (2)$$

A TC with low forward/backward mixing degree has effectively one child/parent, whereas a TC with high forward/backward mixing degree undergoes/results from strong splitting/merging. As shown in SI Sec. V, neither are frequent. It is more common to find weak mixing between TCs, which we believe is due to most papers citing small numbers of papers outside their fields.

At this point, let us recall the Popperian and Kuh-

nian pictures of the evolution of knowledge, where we expect incremental growth punctuated by abrupt paradigm shifts. Certainly, at the aggregate level of PR series of premier physics journals, the number of articles published has grown over the years. When we partition these articles into TCs, we naively expect that some clusters will grow/shrink because of growing/declining interest in their topics. From the alluvial diagrams, we realize that the real picture is far more complex because of recombinations between TCs. Therefore, instead of measuring the growth rates of pure TCs, we need to measure the growth of recombined TCs. To do this, we assume that the contribution of  $C_m^t$  to the size of  $C_n^{t+1}$  is proportional to the size of  $C_m^t$  and also the normalized forward intimacy index  $I_{mn}^f / \sum_n I_{mn}^f$ , i.e.

$$L'(C_n^{t+1}) = \sum_m L(C_m^t) (I_{mn}^f / \sum_n I_{mn}^f). \quad (3)$$

When we plot the predicted sizes  $L'(C_n^{t+1})$  against the observed size  $L(C_n^{t+1})$  in Fig. 3, we find  $(L'(C_n^{t+1}), L(C_n^{t+1}))$  scattered about a straight line with slope with 1.06, which is the annual growth rate of the number of papers in APS journals. This tells us that the growth of recombined TCs is also Popperian.

Next, we consider the Kuhnian processes of splitting and merging. Unlike the “prediction” above, where we made use of information from years  $t$  and year  $t+1$ , we would like to predict the splitting and merging of TCs using information only from year  $t$ . Specifically, for merging events, our ground truth will be the similarity

$$S(C_m^t, C_{m'}^t) = \sum_n (I_{mn}^f / \sum_{n'} I_{mn'}^f) (I_{m'n}^f / \sum_{n''} I_{m'n''}^f) \quad (4)$$

between two TCs in year  $t$ , taking on values between 0 and 1. If  $C_m^t$  and  $C_{m'}^t$  merge perfectly into a single TC in year  $t+1$ ,  $S = 1$ . On the other hand, if the offsprings of

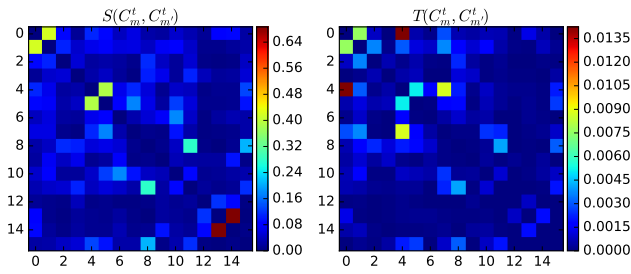


FIG. 4. (left)  $S(C_m^t, C_{m'}^t)$  of 16 TCs in 1991, computed using forward intimacy indices going from 1991 to 1992. (right)  $T(C_m^t, C_{m'}^t)$  of the same 16 TCs, using information from 1991 only. We use the same ordering of TCs in both matrices.

$C_m^t$  and  $C_{m'}^t$  are distinct,  $S = 0$ . To do prediction using only information from year  $t$ , we define

$$T(C_m^t, C_{m'}^t) = W(C_m^t, C_{m'}^t) / (L(C_m^t)L(C_{m'}^t)), \quad (5)$$

where  $W(C_m^t, C_{m'}^t)$  is the sum of weights of edges between papers in  $C_m^t$  and  $C_{m'}^t$ , normalized against the sizes of TCs involved. Fig. 4 shows that the two quantities are highly correlated, with Spearman's rank correlation coefficient of about 0.8 (see SI Sec. VI.) A high  $T$  between two TCs means they are likely to merge the next year.

We also tried to predict the splitting events. Here the situation is more complex: when we examine the weight matrix of a TC, we may find a few large subcommunities or many small subcommunities. Naively, we expect the criterion for splitting is the opposite to merging, i.e. the easier it is to tell one subcommunity from others, the higher the chances for a split. The *boundary index*

$$B = \frac{\sum_{i_1 \neq i_2} \sum_{\substack{j_1 \in C_{i_1} \\ j_2 \in C_{i_2}}} A(j_1, j_2) / \sum_{i_1 \neq i_2} L(C_{i_1})L(C_{i_2})}{\sum_i \sum_{j_1, j_2 \in C_i} A(j_1, j_2) / \sum_i L(C_i)L(C_i)} \quad (6)$$

measures how indistinct the subcommunities are in a TC. Here  $A(j_1, j_2)$  is the weight of the edge between papers  $j_1$  and  $j_2$ , and  $C_i$  is a subcommunity in the given TC. However the picture we find is not as simple as the merging case. When we plot  $M^f$  against  $B$ , we find the expected decreasing trend, but at the same time, the large scatter makes it impossible to reliably predict a splitting event using  $B$ . To better understand the relationship between  $M^f$  and  $B$ , we use quantile regression[20] to find that the  $B$  has no “prediction power” when  $M^f$  is small, but becomes “predictive” when  $M^f$  is large. We summarize these findings in SI Sec. VII. The slopes show that for

the decile of most strongly splitting TCs, increasing the standardized  $B$  by one standard deviation will decrease  $M^f$  by about 0.05, whereas for the decile of the least strongly splitting TCs, there is no obvious trend. In SI Sec. VII we tested another index measuring a different network aspect of the weight matrix and found the prediction results are similar.

Finally, we want to know the impacts of such merging and splitting events. We first check for an increase in

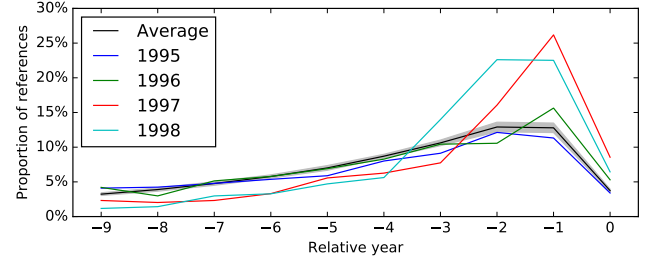


FIG. 5. Proportions of a TC's references published in different years, relative to the year (0) of the TC. The black solid line is the proportions averaged over all TCs in the 1990s, while the area shaded gray is up to one standard deviation away from the mean. Other color lines represent the distribution of four different BEC related TCs.

the number of publications after such events, but found an insignificant difference in paper numbers in strongly and weakly mixing TCs (see Fig. 3(b) and (c)). We suspected this is because our data set is confined to APS publications, and a more careful check should include other physics journals to capture any “influence spillover”. We then check the highest, third quartile, median, average number of citations range two and five years after the events, but still see no significant effects (SI Sec. VIII). Focusing on the highly productive chain of knowledge processes that led to experimental realizations of BEC, quantum teleportation and slow light, we checked the citation profiles between 1995 and 1998. While the 1995 BEC+QI+QO TC cited a slightly lower proportion of 1995 papers than the APS 0-year average, the 1996 BEC+QI+QO, the 1997 BEC TC, the 1998 BEC TCs all cited significantly more 0-year papers. The full effect of this BEC breakthrough can be seen in the large proportions of 1996 papers cited by the 1997 and 1998 TCs and the proportion of 1997 papers cited by the 1998 TC (see Fig. 5). Indeed, we have provided early evidence suggesting that strongly mixing Kuhnian processes are associated with greater impact.

We thank Woo-Sung Jung for discussions.

- 
- [1] K. R. Popper, *All life is problem solving* (Psychology Press, 1999).  
 [2] T. S. Kuhn, *The Structure of Scientific Revolutions (1st ed.)* (University of Chicago Press, 1962).

- [3] D. Wootton, *The invention of science: a new history of the scientific revolution* (Penguin UK, 2015).  
 [4] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, M. A. Rodriguez, and L. Balakireva, PLoS

- ONE **4**, e4803 (2009).
- [5] T. Kuhn, M. Perc, and D. Helbing, *Physical Review X* **4**, 1 (2014), arXiv:1404.3757.
- [6] T. Jia, D. Wang, and B. K. Szymanski, *Nature Human Behaviour* **1**, 0078 (2017).
- [7] P. Chen and S. Redner, *Journal of Informetrics* **4**, 278 (2010), arXiv:0911.0694.
- [8] M. Rosvall and C. T. Bergstrom, *PLoS ONE* **5** (2010), 10.1371/journal.pone.0008694, arXiv:0812.1242.
- [9] D. Chavalarias and J. P. Cointet, *PLoS ONE* **8** (2013), 10.1371/journal.pone.0054847.
- [10] M. Kessler, *IEEE Transactions on Information Theory* **9**, 49 (1963).
- [11] E. Yan and Y. Ding, *Journal of the American Society for Information Science and Technology* **63**, 1313–1326 (2012).
- [12] M. H. Huang, L. Y. Chiang, and D. Z. Chen, *Scientometrics* **58**, 489 (2003).
- [13] “APS Data Sets for Research, see <http://journals.aps.org/datasets>,” .
- [14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008), arXiv:0803.0476.
- [15] M. H. Anderson, J. R. Ensher, M. R. Matthews, C. E. Wieman, and E. A. Cornell, *Science* **269**, 198 (1995).
- [16] K. B. Davis, M. O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, *Physical Review Letters* **75**, 3969 (1995).
- [17] D. Bouwmeester, J.-W. Pan, K. Mattle, M. Eibl, H. Weinfurter, and A. Zeilinger, *Nature* **390**, 575 (1997).
- [18] M. M. Kash, V. A. Sautenkov, A. S. Zibrov, L. Hollberg, G. R. Welch, M. D. Lukin, Y. Rostovtsev, E. S. Fry, and M. O. Scully, *Physical Review Letters* **82**, 5229 (1999).
- [19] L. Jost, *Oikos* **113**, 363 (2006).
- [20] J. Sienkiewicz and E. G. Altmann, *Royal Society Open Science* **3**, 160140 (2016), arXiv:1605.07465.
-

# Knowledge Evolution in Physics Research: An Analysis of Bibliography Coupling Networks

## Supplementary Material

### I. NULL MODEL OF BIBLIOGRAPHY COUPLING NETWORK

To determine the statistical significance of our empirical bibliographic coupling networks (BCNs) (Fig. S1(a)), we build a null model for comparison. In our null model, we fix the out degrees and in degrees of all papers (citing and cited), but rewired the edges to get an ensemble of artificial BCNs (Fig. S1(b)).

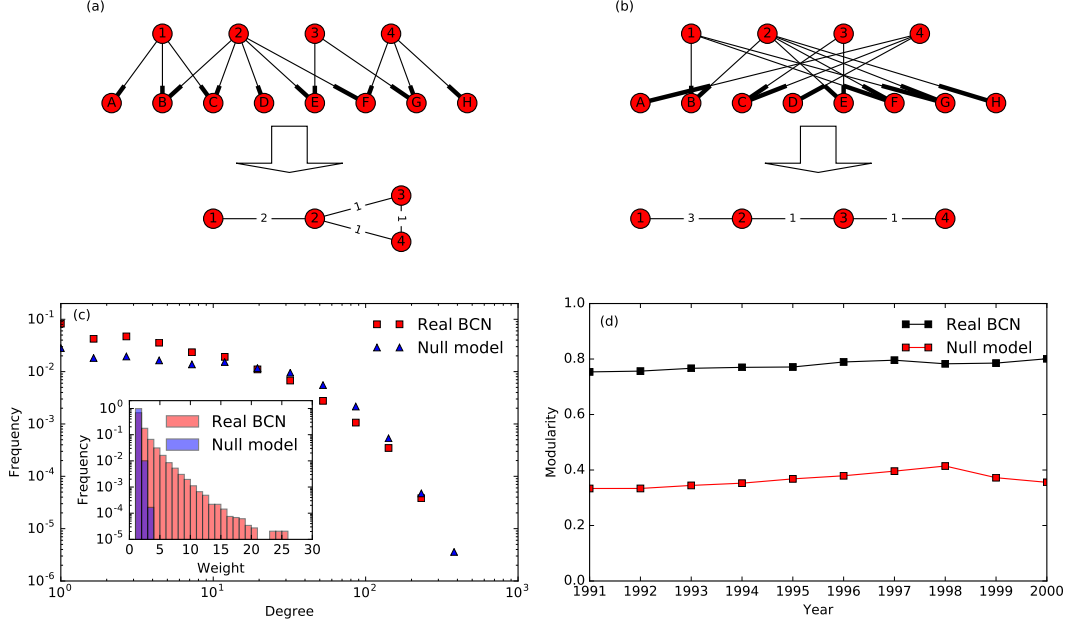


FIG. S1. (a) Original citation network and its BCN. (b) A rewired citation network keeping in degrees and out degrees fixed and its BCN. (c) Comparison of the degree and weight distributions of papers published in 1991, between the real BCN and the null model. (d) Modularities of the best partitions extracted by the Louvain method for the real BCNs and the null model between 1991 and 2000. Results from null model are averaged over 10 different rewirings, and the error bars are much smaller than the marker size.

Compared to the null model, the real BCN has more high-weight edges. We suspect these are the most meaningful edges, arising from the paper's content. If two papers focus on close topics, they will likely have high chance to have more than one common reference, and this effect also manifest itself in the degree distribution: the null model has a flatter degree distribution at small degrees because the edges are drawn by chance, whereas in the real BCN this coupling is based on content, meaning that papers will have edges mostly with papers that are trying to solve the same problems, so the real BCN will have more low-degree nodes, fewer high-degree nodes compared the null model. The most prominent feature of this content-sensitive citation is community structure: in the real BCN, papers focussed on the same topic share more common references with each other than papers focussed on different topics, so that the densities of edges within topics are much higher than between topics. Therefore the modularities of communities extracted by the Louvain method in the real network is much higher than in the null model, as shown in Fig. S1(d).

### II. VALIDATION OF BCN COMMUNITIES

To verify that the communities extracted are really focussed on closely related questions, we check the Physics and Astronomy Classification Scheme (PACS) numbers of members of the communities. Such numbers are provided by authors to indicate the subfields of physics to which their papers belong. In our case, we only use the first two digits of the PACS numbers, as a balance between accuracy and coverage. To test whether the PACS numbers appearing

in the communities could have occurred by chance, we choose one year  $t$ , build its BCN, extracting the community structure with sizes  $\{s_1, s_2, \dots, s_n\}$ , and then randomly assign papers in year  $t$  into  $n$  pseudo-communities of the same sizes, to remove any potential size effects. For a community of size  $s$ , we then identify the largest subset of papers sharing the same PACS number. This PACS number can represent the subfield of the community to a certain extent, and the fraction of papers in the largest subset reflect the homogeneity of the community. On the other hand, the largest subset of papers sharing the same PACS number in a random collection of  $s$  papers is typically small. Dividing the sizes of the largest subsets in the empirical communities and in the random collections, we find ratios are larger than 1 for most cases. This means that the communities we extracted are meaningful (see Fig. S2).

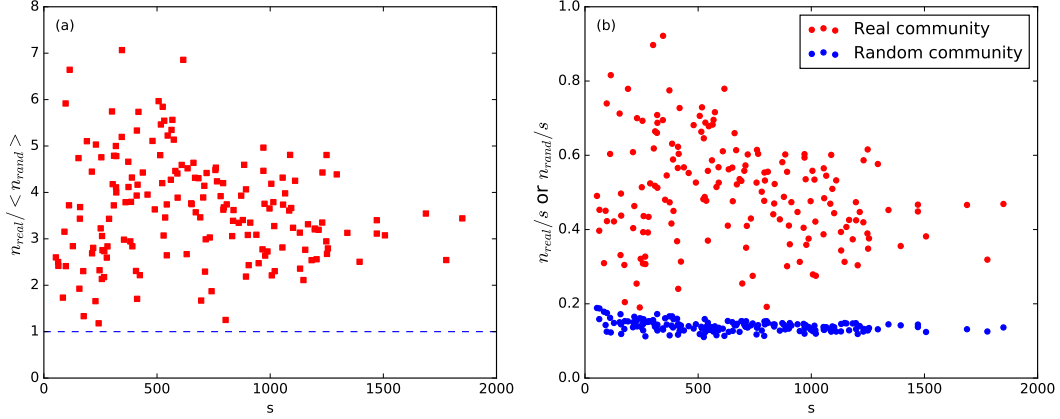


FIG. S2. Comparison of PACS homogeneity between real BCN communities between 1991 and 2000 with more than 50 papers, and their corresponding random collections. (a) The red squares correspond to the sizes of the largest subsets of papers sharing at least one PACS number,  $n_{real}$ , in the empirical communities divided by the same quantity found in the corresponding random collections,  $n_{rand}$ , as a function of the community size  $s$ . (b) The fraction of the largest subset of papers sharing at least one PACS number as a function of  $s$  for real communities in the BCN and random collections. For clarity, the small error bars are not shown in the figures.

### III. ALLUVIAL DIAGRAM FOR 1991-2000

In addition to Fig. 2 in the letter, we also plotted an alluvial diagram for 1991 to 2000, showing the splitting of PRA into PRA and PRE. As we can see from Fig. S3, before 1993, there were several PRA-dominated TCs. After the split in 1993, some PRA-dominated TCs remained PRA-dominated, whereas other PRA-dominated TCs became PRE-dominated. This means that even before 1993, papers in PRA were already divided into groups based on different topics, some of which are predecessors of the PRE TCs.

### IV. CASE STUDY: QUANTUM OPTICS, QUANTUM INFORMATION AND BOSE-EINSTEIN CONDENSATION

To illustrate the utility our knowledge evolution framework can offer, we use as a case study the interesting interactions between quantum optics (QO), quantum information (QI), and Bose-Einstein Condensation (BEC). These three fields experienced breakthroughs in the 1990s. In Fig. S4 we highlight the evolution of TCs which are related to these three topics and Table S1 shows the three most cited papers in these TCs. Key merging and splitting events are reported in the main paper, as are important publications these events are correlated with.

### V. KNOWLEDGE METABOLISM

Some TCs have more references overlapping with those in the previous year, while other TCs have less. To quantify evolution of references, we count the sums of the forward and backward intimacy indices. These represent the percentage of references going to the next year, and the percentage of references inherited from the previous year, which we think of as the ‘outflow’ and ‘inflow’ respectively. As shown in Fig. S5(a) and (b), most outflows and inflows



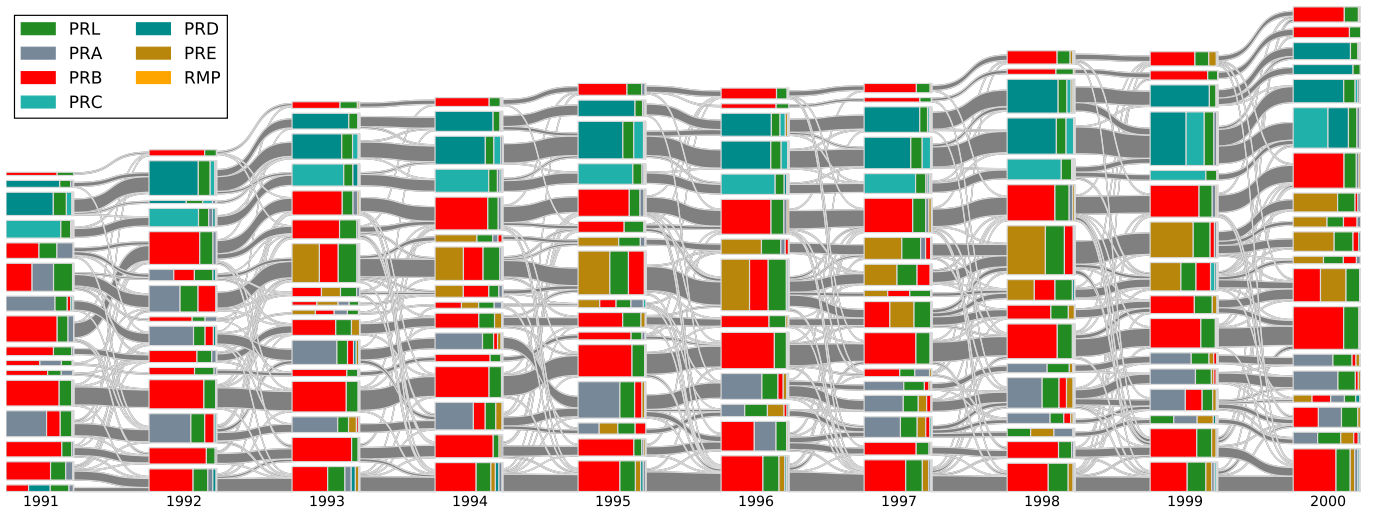


FIG. S3. The alluvial diagram of APS papers from 1991 to 2000. Each block in a column represents a TC and the height of the block is proportional to the number of papers in the TC. Only communities comprising more than 100 papers are shown. TCs in successive years are connected by streams whose widths at the left and right ends are proportional to the forward and backward intimacy indices. The different colors in a TC represent the relative contributions from different journals.

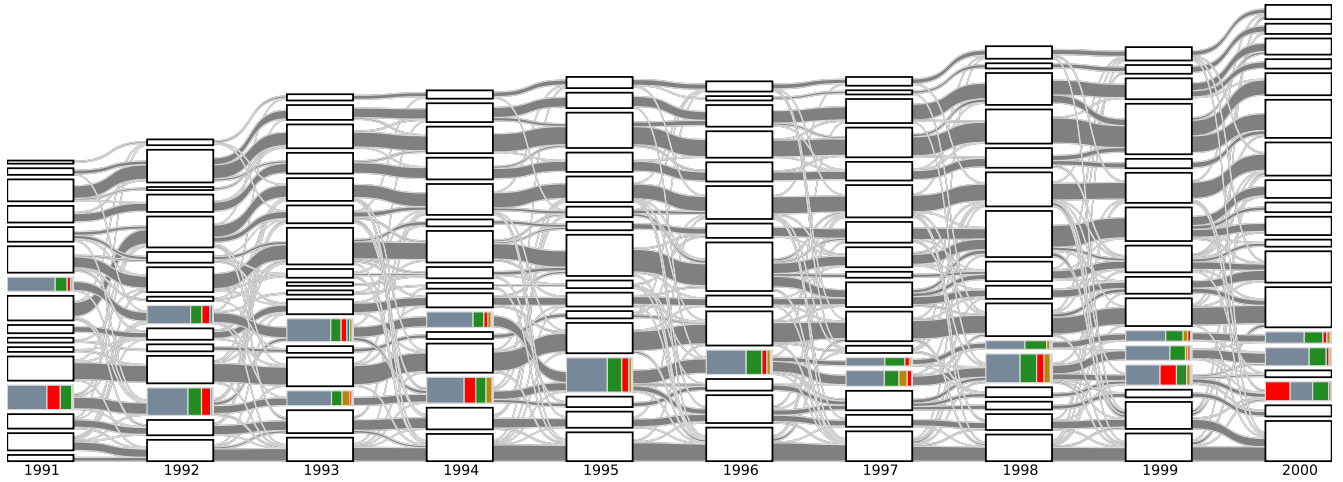


FIG. S4. The alluvial diagram of APS papers from 1991 to 2000, where we colored only TCs highly related to quantum optics, quantum information and Bose-Einstein condensation.

are distributed within a narrow range, but there are exceptional cases as well: such as a single peak in Fig. S5(b), whose references overlap significantly less than normal with the previous year. In the context of birth, death, growth, decay, split, and merge knowledge processes, we are inclined to call this event in 1993 the birth of a TC. Further analysis shows that most common PACS codes are: 03 (Quantum mechanics, field theories, and special relativity), 42 (Optics) and 63 (Lattice dynamics). Looking at the references of this TC, we find that most of these comes from 1990, 3 year before. This interesting phenomenon is therefore more appropriately identified as a sleeping beauty[S1].

Every year, physicists absorb new references and drop old references as their fields progress. Although this ‘metabolism’ differ from TC to TC, the whole process is quite stable over all TCs, as shown in Fig. S5(c) and (d). This universal curve can be used as a benchmark for the test of scientific impact, as we have done in Fig. 5 of the letter.

As we see from Fig. S3, there is a great diversity of processes acting on the TCs: some TCs are derived almost exclusively from one source, others receive strong contributions from a small number of sources, or weak contributions from a large number of sources. To quantify such diversity, we introduce forward mixing degree and backward mixing degree (Eq. 2 in the letter) As shown in Fig. S5(e), (f), strong splitting/merging or almost isolated development are



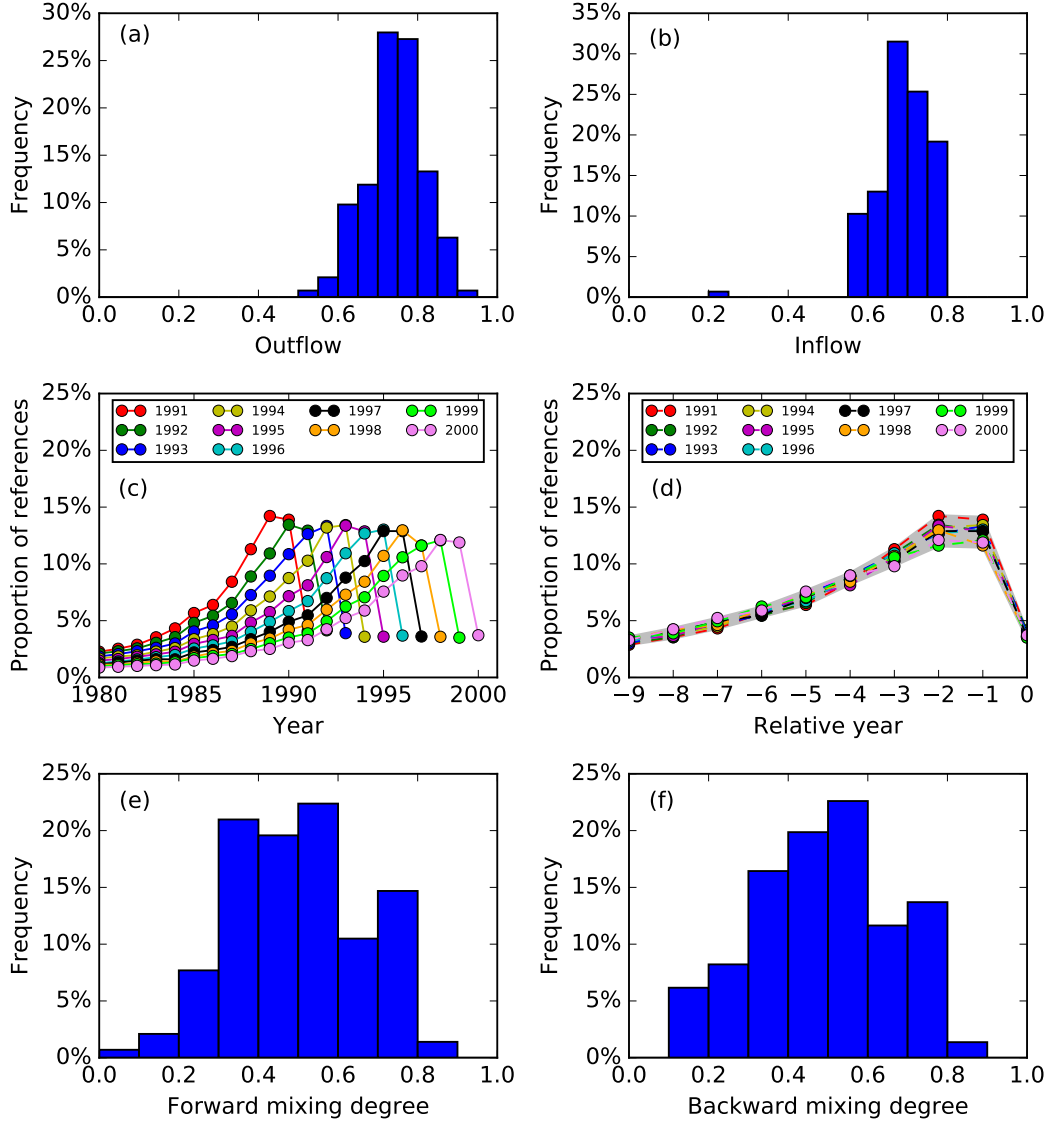


FIG. S5. The metabolic analysis of APS papers in the 1990s. (a) The distribution of outflows of TCs. (b) The distribution of inflows of TCs. (c) Proportions of APS paper's references published in different years. (d) Proportions of APS paper's references published in different years, relative of the year (0) of publication. (e) The distribution of forward mixing degree of TCs. (f) The distribution of backward mixing degree of TCs.

rare, and in most case, TCs undergo weak information exchange.

## VI. PREDICTION OF MERGING

As mentioned in the letter, we found that the inter-TC connection is highly correlated with their mixing the next year. High  $T(C_m^t, C_{m'}^t)$  leads with a large probability to a high  $S(C_m^t, C_{m'}^t)$ . Analyzing APS papers in the 1990s, we found a Spearman's rank coefficient of 0.804 between  $T(C_m^t, C_{m'}^t)$  and  $S(C_m^t, C_{m'}^t)$  over all TCs (with at least 100 papers). However, because the average Pearson correlation coefficient is only 0.504, such a relation is not linear (see Fig. S6)

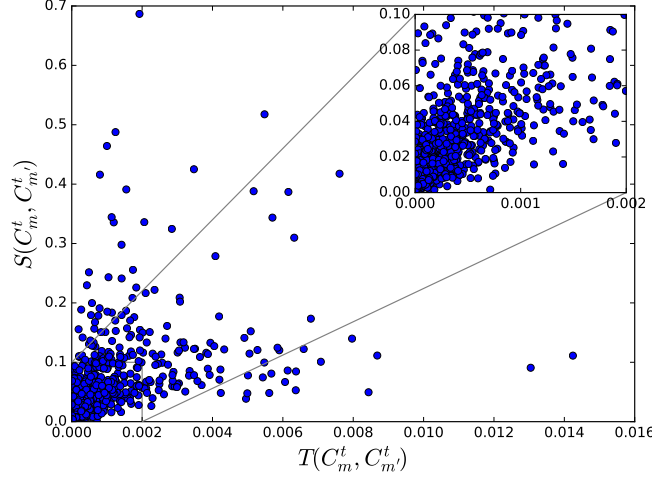


FIG. S6. The scatter plot between  $T(C_m^t, C_{m'}^t)$  and  $S(C_m^t, C_{m'}^t)$  among all TCs (with at least 100 papers) in 1990s.

## VII. HIERARCHY STRUCTURE OF COMMUNITY AND SPLITTING ANALYSIS

To predict a splitting event, we check the weight matrix of different TCs. Naively, we expect the components that lead on to different TCs the next year to already form distinct subcommunities this year. However, when we use the dendrogram extracted from the Louvain method to identify subcommunities, we found that different TCs have different internal structures (see Fig. S7): some have a few large subcommunities, while others have many small subcommunities.

Assuming that splitting is the time reversal of merging, i.e. two distinct subcommunities becoming two distinct communities in the next year, we devise the boundary index  $B$  to measure how indistinct the subcommunities are in a TC. Quantile regression (QR) shows that the relation between  $B$  and  $M^f$  depends on the decile, as shown in Fig. S8(a), (b). For different decile, the relation is different.

We also try with *fragmentation index*

$$F = \sum_{i:j[i]} w_i s_{j[i]}^2 \quad (\text{S1})$$

where  $w_i$  is the size fraction of the top level subcommunity  $i$ ,  $s_{j[i]}$  is the relative size fraction of subsubcommunity  $j$  inside subcommunity  $i$ . The more fragmentation a community is, the closer  $F$  is to 0. Quantile regression between  $F$  and  $M^f$  gives very similar results as  $B$  and  $M^f$ , that is, for the decile of most strongly splitting TCs, increasing the standardized  $F$  by one standard deviation will decrease  $M^f$  by about 0.06, whereas for the decile of the least strongly splitting TCs, there is no obvious trend as  $\beta$  close to 0, as shown in Fig. S8(c), (d).

## VIII. CORRELATION ANALYSIS BETWEEN CITATION AND MIXING

When we think of high-impact research, we think of highly-cited papers. Therefore, to quantify the impact of strongly-splitting events in the alluvial diagrams, we counted the citations of TCs resulting from splittings. As shown in Fig. S9, we did this for number of citations 2 years after the events, and also 5 years after the events. There were no obvious trends. The results of backward mixing degree, i.e. merging, are similar.

---

[S1] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, Proceedings of the National Academy of Sciences of the United States of America **2015**, 40 (2015), arXiv:1505.06454.

TABLE S1. The three most cited papers in quantum optics, quantum information theory, quantum computation and Bose-Einstein condensation related TCs.

Year	TC	DOI	Title
1991	Upper	10.1103/PhysRevLett.67.661	Quantum cryptography based on Bells theorem
		10.1103/PhysRevLett.66.2593	Observation of electromagnetically induced transparency
		10.1103/PhysRevLett.67.1855	Enhancement of the index of refraction via quantum coherence
	Lower	10.1103/PhysRevA.44.5674	Above-surface neutralization of highly charged ions: The classical over-the-barrier model
1992	Upper	10.1103/PhysRevB.43.13401	Strong magnetic x-ray dichroism in 2p absorption spectra of 3d transition-metal ions
		10.1103/PhysRevLett.66.2601	Dynamic stabilization of hydrogen in an intense, high-frequency, pulsed laser field
		10.1103/PhysRevLett.69.2881	Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states
	Lower	10.1103/PhysRevLett.69.3314	Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity
		10.1103/PhysRevLett.68.580	Wave-function approach to dissipative processes in quantum optics
		10.1103/PhysRevLett.68.1943	X-ray circular dichroism as a probe of orbital magnetization
1993	Upper	10.1103/PhysRevLett.68.3535	High-order harmonic generation from atoms and ions in the high intensity regime
		10.1103/PhysRevLett.69.1383	Absorption of ultra-intense laser pulses
		10.1103/PhysRevLett.70.1895	Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels
	Lower	10.1103/PhysRevA.47.4114	Threshold and resonance phenomena in ultracold ground-state collisions
		10.1103/PhysRevLett.70.1244	Measurement of the Wigner distribution and the density matrix of a light mode using optical homodyne tomography: Application to squeezed states and the vacuum
		10.1103/PhysRevLett.71.1994	Plasma perspective on strong field multiphoton ionization
1994	Upper	10.1103/PhysRevLett.70.1599	Above threshold ionization beyond the high harmonic cutoff
		10.1103/PhysRevLett.70.774	High-order harmonic generation in rare gases with a 1-ps 1053-nm laser
		10.1103/PhysRevA.50.67	Squeezed atomic states and projection noise in spectroscopy
	Lower	10.1103/PhysRevLett.72.3439	Statistical distance and the geometry of quantum states
		10.1103/PhysRevLett.73.58	Experimental realization of any discrete unitary operator
		10.1103/PhysRevA.49.2117	Theory of high-harmonic generation by low-frequency laser fields
1995	Upper	10.1103/PhysRevLett.73.1227	Precision Measurement of Strong Field Double Ionization of Helium
		10.1103/PhysRevA.50.1540	Modeling harmonic generation by a zero-range potential
		10.1103/PhysRevLett.75.3969	Bose-Einstein Condensation in a Gas of Sodium Atoms
1996	Upper	10.1103/PhysRevLett.74.4091	Quantum Computations with Cold Trapped Ions
		10.1103/PhysRevA.52.R2493	Scheme for reducing decoherence in quantum computer memory
		10.1103/PhysRevA.54.3824	Mixed-state entanglement and quantum error correction
1997	Upper	10.1103/PhysRevLett.77.1413	Separability Criterion for Density Matrices
		10.1103/PhysRevLett.77.2360	Collective Excitations of a Trapped Bose-Condensed Gas
		10.1103/PhysRevLett.78.985	Bose-Einstein Condensation of Lithium: Observation of Limited Condensate Number
	Lower	10.1103/PhysRevLett.78.586	Production of Two Overlapping Bose-Einstein Condensates by Sympathetic Cooling
		10.1103/PhysRevLett.78.5	Demonstration of the Casimir Force in the 0.6 to 6 $\mu$ m Range
		10.1103/PhysRevLett.78.5022	Entanglement of a Pair of Quantum Bits
1998	Upper	10.1103/PhysRevLett.78.3221	Quantum State Transfer and Entanglement Distribution among Distant Nodes in a Quantum Network
		10.1103/PhysRevLett.79.3306	Noiseless Quantum Codes
		10.1103/PhysRevLett.81.3108	Cold Bosonic Atoms in Optical Lattices
	Lower	10.1103/PhysRevLett.81.938	Atomic Scattering in the Presence of an External Confinement and a Gas of Impenetrable Bosons
		10.1103/PhysRevLett.81.742	Spinor Bose Condensates in Optical Traps
		10.1103/PhysRevA.57.120	Quantum computation with quantum dots
1999	Upper	10.1103/PhysRevLett.80.2245	Entanglement of Formation of an Arbitrary State of Two Qubits
		10.1103/PhysRevLett.81.5932	Quantum Repeaters: The Role of Imperfect Local Operations in Quantum Communication
		10.1103/PhysRevLett.83.2498	Vortices in a Bose-Einstein Condensate
	Middle	10.1103/PhysRevLett.83.5198	Dark Solitons in Bose-Einstein Condensates
		10.1103/PhysRevLett.82.1975	Entanglement of Atoms via Cold Controlled Collisions
		10.1103/PhysRevLett.83.4204	Quantum Information Processing Using Quantum Dot Spins and Cavity QED
2000	Upper	10.1103/PhysRevB.59.2070	Coupled quantum dots as quantum gates
		10.1103/PhysRevLett.82.2417	Dynamical Decoupling of Open Quantum Systems
		10.1103/PhysRevLett.82.5229	Ultraslow Group Velocity and Enhanced Nonlinear Optical Effects in a Coherently Driven Hot Atomic Gas
	Lower	10.1103/PhysRevLett.83.2845	Transmission Resonances on Metallic Gratings with Very Narrow Slits
		10.1103/PhysRevLett.83.967	Liquid-Crystal Photonic-Band-Gap Materials: The Tunable Electromagnetic Vacuum
		10.1103/PhysRevLett.84.806	Vortex Formation in a Stirred Bose-Einstein Condensate
2001	Upper	10.1103/PhysRevLett.85.1795	Stable 85Rb Bose-Einstein Condensates with Widely Tunable Interactions
		10.1103/PhysRevLett.85.3745	Regimes of Quantum Degeneracy in Trapped 1D Gases
		10.1103/PhysRevA.62.062314	Three qubits can be entangled in two inequivalent ways
	Middle	10.1103/PhysRevLett.84.2722	Inseparability Criterion for Continuous Variable Systems
		10.1103/PhysRevA.62.012306	Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures
		10.1103/PhysRevLett.85.5214	Double Resonant Raman Scattering in Graphite
2002	Lower	10.1103/PhysRevLett.85.154	Electronic Structure of Deformed Carbon Nanotubes
		10.1103/PhysRevB.62.13104	Carbon nanotubes, buckyballs, ropes, and a universal graphitic potential

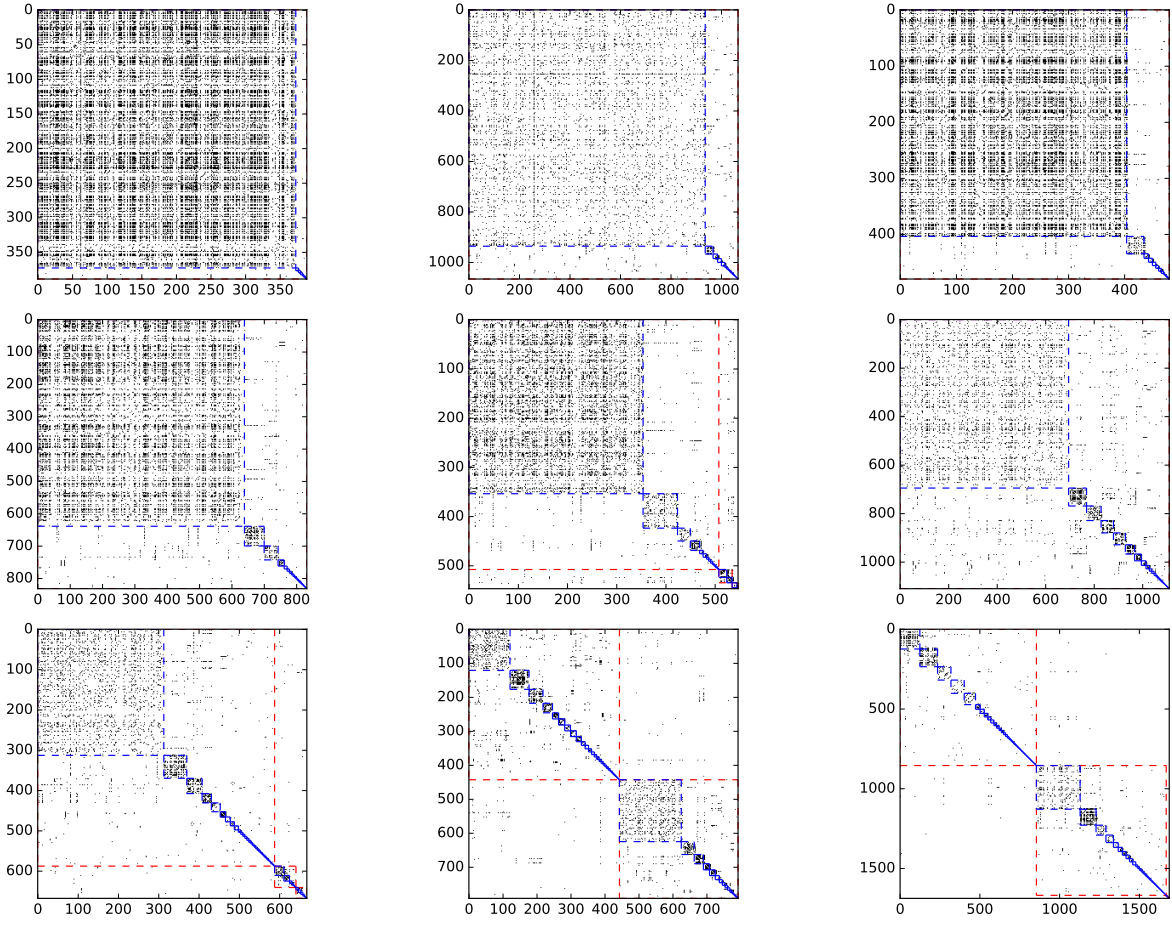


FIG. S7. Adjacency matrices of TCs in the 1990s. The blue lines indicate the boundaries of subsubcommunities, the red lines indicate the boundaries of subcommunities. The red lines are absent from some plots because such TC have only one level when the Louvain algorithm terminated.

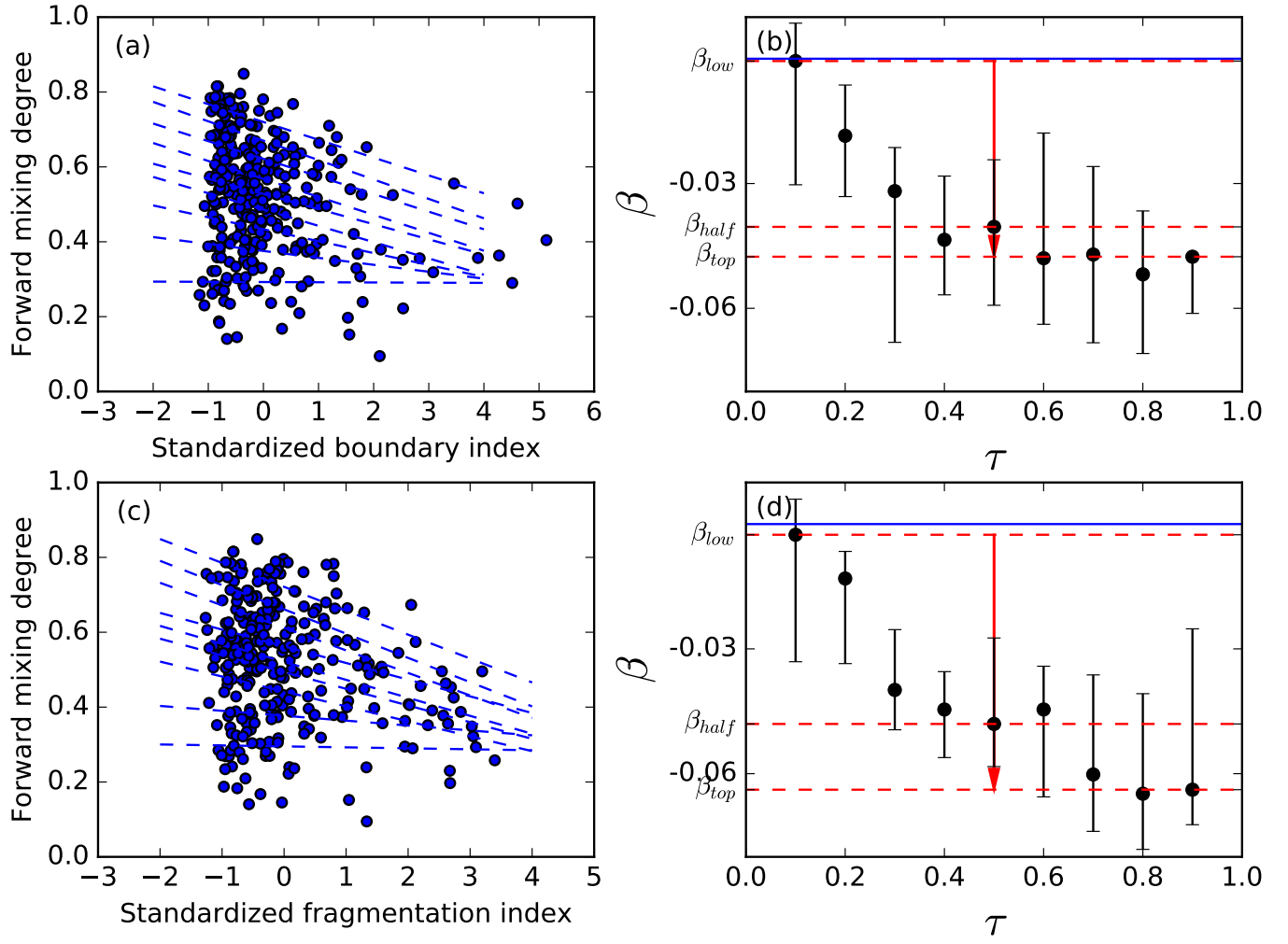


FIG. S8. Relation between boundary index, fragmentation index and forward mixing degree of TCs in 1980s and 1990s. (a) Each dot corresponds to one TC, dash lines show QR results for quantiles  $\tau = 0.1, 0.2, \dots, 0.9$ . (b)  $\beta$  coefficients (slopes of QR in the (a)) as a function of  $\tau$ . The red arrows show  $\beta_{low} \equiv \beta(\tau = 0.1)$ ,  $\beta_{half} \equiv \beta(\tau = 0.5)$  and  $\beta_{top} \equiv \beta(\tau = 0.9)$ , as, respectively, the nock, a circle on the shaft, and the head of the arrow, the blue solid line represents 0. (c) Each dot corresponds to one TC, dash lines show QR results for quantiles  $\tau = 0.1, 0.2, \dots, 0.9$ . (d)  $\beta$  coefficients (slopes of QR in the (c)) as a function of  $\tau$ . The red arrows show  $\beta_{low} \equiv \beta(\tau = 0.1)$ ,  $\beta_{half} \equiv \beta(\tau = 0.5)$  and  $\beta_{top} \equiv \beta(\tau = 0.9)$ , as, respectively, the nock, a circle on the shaft, and the head of the arrow, the blue solid line represents 0.

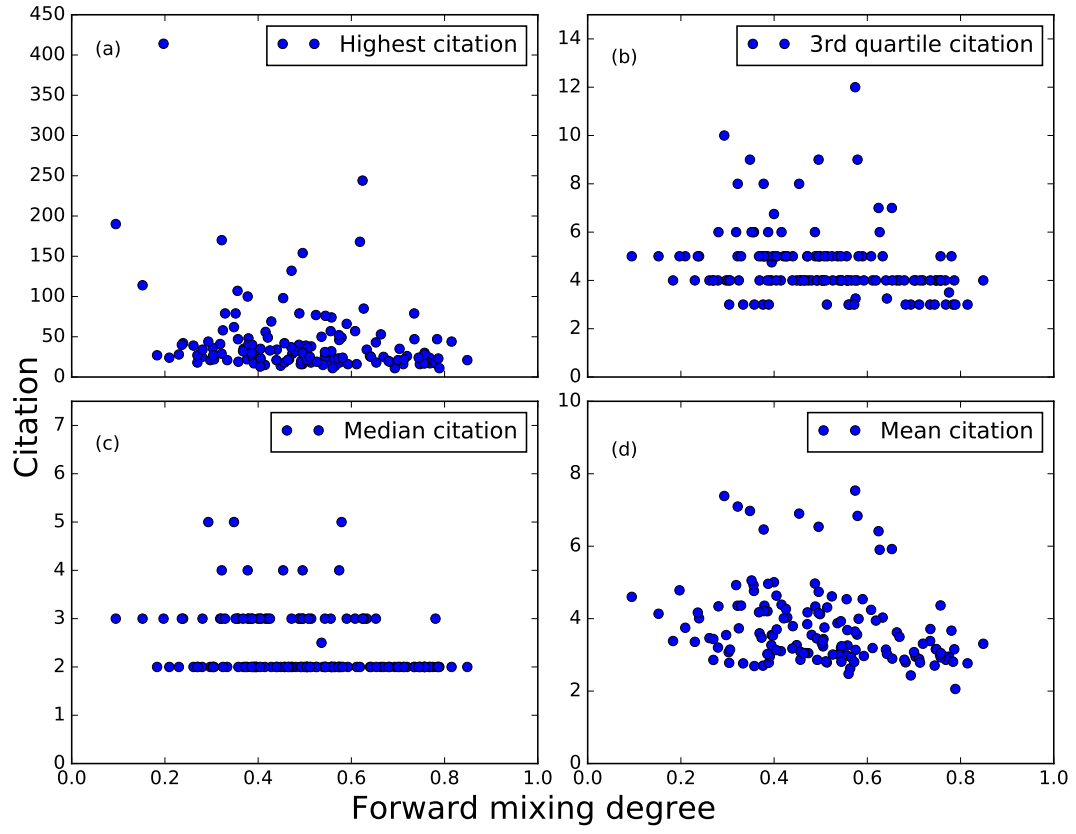


FIG. S9. The scatter plot between different citations received during 2 years and forward mixing degree among all TCs (with at least 100 papers) in 1990s. (a) Highest citation, (b) Third quartile citation, (c) Median citation, (d) Mean citation.